The Band-Gap Problem in Semiconductors Revisited: Effects of Core States and Many-Body Self-Consistency

Wei Ku* and Adolfo G. Eguiluz

Department of Physics and Astronomy, The University of Tennessee, Knoxville, TN 37996-1200, and Solid State Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831–6030

A novel picture of the quasiparticle (QP) gap in prototype semiconductors Si and Ge emerges from an analysis based on all-electron, self-consistent, GW calculations. The deep-core electrons are shown to play a key role via the exchange diagram —if this effect is neglected, Si becomes a semimetal. Contrary to current lore, the Ge $3d$ semicore states (e.g., their polarization) have no impact on the GW gap. Self-consistency improves the calculated gaps —a first clear-cut success story for the Baym-Kadanoff method in the study of real-materials spectroscopy; it also has a significant impact on the QP lifetimes. Our results embody a new paradigm for *ab initio* QP theory.

The modern "band-gap problem" originated with the realization that density-functional theory (DFT) [1], implemented in the local-density approximation (LDA), failed drastically in the description of the fundamental excitation gap of semiconductors and insulators. A significant step forward was achieved in the mid-eighties, when the first *ab initio* calculations of quasiparticle (QP) states were performed [2, 3] within Hedin's GW approximation (GWA) [4]. At the present time, it is nearly-universally accepted [2, 3, 4, 5, 6, 7, 8, 9, 10] that the GWA yields QP gaps in semiconductors and insulators to within 0.1 eV of experiment —which is the level of accuracy required in the study of transport in these materials [6].

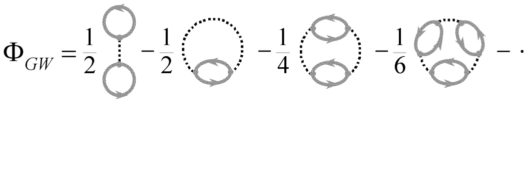
In this Letter we uncover a novel picture of the physical ingredients underlying the observed QP gap in Si and Ge. Central elements of this picture are the impact of the core electrons on the many-body problem for the states at the gap, and the role of self-consistency. The GW schemes alluded to above turn out to benefit from "cancellation of errors" involving the neglect of both effects. Our results illustrate the practical importance of the Baym-Kadanoff conserving method [11] for the study of excitations in real materials.

These conclusions are arrived at by eliminating approximations which are routinely introduced in the implementation of Hedin's scheme. First, the usual GW work invokes the pseudopotential (PS) approximation — by which the core states are effectively eliminated from the gap problem. However, PS theory does not guarantee that a "partitioning" of the electrons into two groups may lead to an accurate description of the dynamical self-energy of the valence states —which, according to DFT, is a non-linear functional of the *total* density. Semicore states pose a special challenge [6, 7]; significantly, on the basis of a phenomenological model, the indirect nature of the Ge gap has been assigned to an effect of the polarization of the $3d$ states [6].

Second, in most GW calculations the Dyson equation (DEq) is not solved to self-consistency. However, it has been shown, for a Hubbard-type model, that this practice leads to a genuine violation of charge conservation [12]. Still, from the available self-consistent solutions of the DEq [14, 15, 16] it has been inferred that, while self-consistency, at the GW level, is a must in total-energy calculations [17], the same is "to be avoided" in the study of spectroscopy [5, 14, 15, 18]. This state of affairs is unsatisfactory, since self-consistency is a necessary condition for the fulfillment of all the conservation laws [11] —and, thus, for a proper theory of transport [13].

In our GW calculations all the electrons are taken into account in the evaluation of the valence-electron self-energy. Remarkably, the deep core states are found to play a significant role in the QP gap problem via the core-valence exchange diagram. An additional surprise is that the (presumably important) shallow Ge $3d$ semicore states have no effect on the GW gap. Self-consistency at the GW level does improve the QP gaps; it also impacts the QP lifetimes. Our results for the Si gap and the indirect Ge gap al L agree with experiment very well. Other aspects of our calculated QP band structures provide signatures of physics beyond the GWA.

We recall that the exact self-energy is "Φ-derivable," [19] $\Sigma\left[G\right](11') = \delta\Phi/\delta G\,(1'1)$, where Φ is the Luttinger-Ward "free-energy" functional; our notation stresses the fact that Σ is a functional of the dressed Green's function G. Now, a Φ-derivable Σ, obtained on the basis of an *approximate* Φ-functional, coupled with a self-consistent solution of the DEq, $G^{-1}\left(1,1'\right) = G_0^{-1}\left(1,1'\right) - \left(\Sigma\left[G\right]\left(1,1'\right) - \left(\mu-\mu_0\right)\delta\left(1-1'\right)\right)$, ensures that G fulfills the conservation laws *exactly* [11]. [Here $1,1'$ denote space-time points; Matsubara times τ are defined for $0 \le \tau \le \beta\hbar$. μ, the chemical potential for the correlated system with a fixed number of electrons, is obtained self-consistently with G; μ differs appreciably ($\sim$1eV) [20] from its counterpart μ_0 for the reference one-electron system whose Green's function is G_0.] The GWA [4] is defined by the Φ-functional



, (1)

where the particle-hole bubbles are made up of G's (not G_0's), and the dashes represent the Coulomb interaction v. Functional differentiation of Φ_{GW} yields $\Sigma_{GW}[G] = \Sigma_H[G] + \Sigma_{xc}[G]$, where $\Sigma_H[G]$ is the Hartree term, and the exchange-correlation (XC) term is of the Hedin form $\Sigma_{xc}[G](1,1') = -G(1,1')\,W[G](1,1')$, where $W[G](1,1')$ is the screened interaction [4].

We work in the basis of the Kohn-Sham (KS) states $\phi_{k,j}(x)$ provided by the full-potential, linearized augmented-plane-wave (FLAPW) method [21]; here k is a wave vector in the Brillouin zone, and j is a band index. Adopting (without lack of generality) the KS system as the reference one-particle system, described within the LDA, the DEq can be written as [22]

$$G_{k,j}(\tau-\tau') = G^{LDA}_{k,j}(\tau-\tau') + G^{LDA}_{k,j}(\tau-\bar{\tau}_1) \\ \times \left[\Sigma_{k,j}(\bar{\tau}_1-\bar{\tau}_2) - \left(V^{LDA}_{k,j} + (\mu-\mu_0)\right)\delta(\bar{\tau}_1-\bar{\tau}_2)\right] \\ \times G_{k,j}(\bar{\tau}_2-\tau'), \quad (2)$$

where $G^{LDA}_{k,j}(\tau) = -\frac{1}{\hbar}e^{-\varepsilon_{k,j}\tau/\hbar}\left(\theta(\tau) - n_F(\varepsilon_{k,j})\right)$, the KS eigenvalues $\varepsilon_{k,j}$ being measured from μ_0, V^{LDA} is the KS potential without the nuclear contribution, and summation over variables with a bar on top is understood.

We solve Eq. (2) and the integral equation for $W[G](1,1')$ (in the latter case, in reciprocal space) on the τ–axis. Our approach is ideally suited for a self-consistent evaluation of G and should prove valuable in calculations beyond the GWA [20]. A novel feature of our scheme —which is devoid of the cutoff effects encountered in ω-axis formulations [16]— is the use of a non-uniform "power mesh" (PM) [20] which, as outlined in Fig. 1, accounts for the nature of both G and the particle-hole bubble —which are strongly peaked at the ends of the interval $0 \le \tau \le \beta\hbar$, being flat in between; moreover, the PM allows us to perform high-order interpolation (scaling ~linearly) to generate all functions on the dense, uniform mesh required in the evaluation of the products entering the τ-integrals [20]. From the solution of Eq. (2) we evaluate the spectral function $A_{k,j}(\omega) = -\frac{1}{\pi}\text{Im}G_{k,j}(\omega)$ via analytic continuation of G onto the real-ω axis [23]. For the j-bands of interest, $A_{k,j}(\omega)$ shows a well-defined peak, whose $k-\omega$ dependence yields the QP band structure —see Fig. 1.

The convergence of our results was tested by varying the parameters involved in: $\boldsymbol{k}$-space sampling (we used 5x5x5 and 8x8x8 meshes), number of bands (14 and 24), number of (reciprocal lattice-) $\boldsymbol{K}$-vectors used in the evaluation of Σ_{xc} (9, 27, 51, and 65), temperature (4000, 2000, 1000, and 300 K), and PM mesh (p5u20, p6u10, and p6u20). The most demanding parameter, the number of $\boldsymbol{K}$-vectors kept in the valence exchange term, is associated with a monotonic opening of the Si gap. (The core contribution converges even slower; thus, a 6D real-space integral is evaluated instead.) In the case of Si we estimate that the absolute gap is converged to ~0.1 eV from below; for Ge the convergence is even better, from both directions.

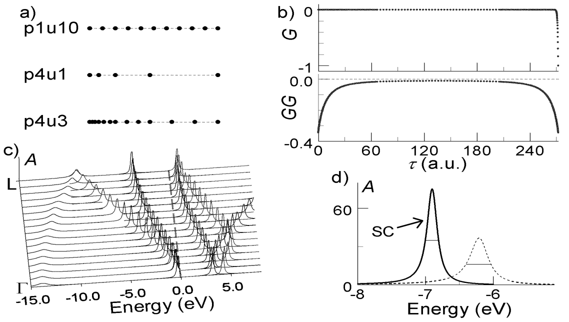


FIG. 1: a): The PM used to solve the DEq on the τ-axis is defined by two integers: "p" is the "order" of the underlying non-uniform mesh, whose width doubles in each step; "u" is the number of uniform intervals into which the non-uniform intervals are partitioned. b): Typical τ-dependence of G and of the particle-hole bubble (GG); their exponential localization (and discontinuity) at $\tau = 0$ and $\beta\hbar$ is efficiently accounted for by our PM [20]. c): Spectral function $A_{k,j}(\omega)$ for Si along ΓL; to aid visualization, small numerical broadening has been introduced. d): $A_{k,j}(\omega)$ for the hole state at the midpoint of the occupied band; the solid/dashed line denotes the self-consistent/first-iteration solution of Eq. (2).

Table I provides the framework for a discussion of our results for the absolute QP gap (located at ~80% ΓX), the direct gap at Γ, and the occupied bandwidth of Si. Our results comprise two levels of implementations of the GWA: (*i*) A self-consistent solution of Eq. (2), corresponding to the evaluation of the self-energy $\Sigma_{GW}[G_{GW}]$ (3^{rd} row); *this calculation represents a numerical realization of a conserving approximation* [11]. *(ii)* A non-conserving calculation (4^{th} row) corresponding to the use of the self-energy $\Sigma_{GW}[G_{LDA}]$ —the standard "$G_{LDA}\,W_{LDA}$" approximation [2, 3, 4, 5, 6, 7, 8, 9, 10]. Clearly, the gaps obtained from $\Sigma_{GW}[G_{GW}]$ are in good agreement with experiment [24]; the mechanisms behind this agreement turn out to be quite unexpected —and instructive.

Indeed, on the basis of several numerical tests, we uncovered, first of all, the role of the deep-core electrons. In one set of calculations we suppressed their contribution to the exchange self-energy —2^{nd} term in Eq. (1)— for the states at the gap [25]. At the $\Sigma_{GW}[G_{LDA}]$ level, the absolute gap (~0.85 eV, see Table I) is then reduced by 0.9 eV —*i.e., the gap is closed, the ensuing QP band structure of Si corresponding to a semi-metal.* While the size of this effect is surprising, its physics is easy to visualize: (*i*)The core electrons shield the attractive field of the nuclei (via the 1^{st} term in Eq. (1), thereby raising the energy of the valence and conduction states; *(ii)* the exchange process partially compensates for this effect; *(iii)* the states across the gap have different amplitudes in the core region; these amplitudes control the strength of the

TABLE I: QP band gaps and occupied bandwidth of Si. Comparison of our *all-electron* GW results with experiment, with the (approximate) all-electron GW calculations of Refs. [9] and [10], and with a representative PS-based GW calculation. The 3^{rd} row obtains from our fully conserving self-energy $\Sigma_{GW}\,[G_{GW}]$; the 4^{th} row, and all rows below it, obtain from the non-conserving approximation $\Sigma_{GW}\,[G_{LDA}]$.

	Absolute gap	Direct gap at Γ	Occupied bandwidth
Experiment [24]	1.17	3.40	12.5 ± 0.6
LDA (FLAPW)	0.52	2.53	12.22
$\Sigma_{GW}\,[G_{GW}]$	1.03	3.48	13.53
$\Sigma_{GW}\,[G_{LDA}]$	0.85	3.12	12.15
∼all-electron [9]	1.01	3.30	12.21
∼all-electron [10]	1.00	3.15	. . .
PS-based [2]	1.29	3.35	12.04

exchange. It is the larger lowering of the energy of the QP states below the gap, relative to those above it, *due to the non-local core-valence exchange process*, that leads to this novel all-electron effect [26].

Further insight into the role of the core electrons is obtained by simulating their contribution to $\Sigma_{xc}\,[G_{LDA}]$ as $\Sigma_{xc}^{\text{from core}} \approx V_{xc}^{LDA}\,[n_{tot}] - V_{xc}^{LDA}\,[n_{val}]$, where n_{tot} and n_{val} are the total and valence densities, and V_{xc}^{LDA} is the XC contribution to V^{LDA}. This uncontrolled "LDA recipe" yields a *spurious* 0.15 eV additional opening of the Si gap [20]. The significance of this test is that it accounts (Table I) for the difference between the *approximate* FLAPW-based $\Sigma_{GW}\,[G_{LDA}]$-level result of Hamada, Hwang, and Freeman [9] (1.01 eV) and our own (0.85 eV) —Ref. [9] relies on this LDA *ansatz* for the core contribution to the valence-electron self-energy, as the PS-based GW schemes implicitly do.

From Table I we draw a second key message: contrary to current wisdom [4, 14, 15, 16, 18], self-consistency does improve the quality of the calculated GW gaps of Si; cf. rows 1, 3, and 4. Indeed, the additional opening of the gaps obtained from $\Sigma_{GW}\,[G_{GW}]$ brings them closer to their experimental values. This effect is traced to the dressing of both W and G. Indeed, the dynamical screening built into W, which greatly reduces the exaggerated Hartree-Fock (HF) gap, is weakened for the dressed W; the dressing of G widens the gap as well, a trend most easily visualized within self-consistent HF [20]. We stress that *the success of our conserving scheme is intimately related to the fact that we have carried out a full all-electron calculation* [27].

Interestingly, self-consistency also has a significant effect on the QP *lifetimes*, given by the inverse of the full-width at half-maximum (FWHM) of the QP peak in $A_{\vec{k},j}\,(\omega)$. As seen in Fig. 1d) for the hole state at the midpoint of the Si band, the FWHM (solid line) is significantly reduced, relative to its non-self-consistent counterpart (dashes). This effect —which has been ignored in the rapidly-growing literature on "hot carrier" lifetimes [28]— recognizes two sources. First, fully dressed QP's scatter less frequently off the Fermi sea than in the $\Sigma_{GW}\,[G_{LDA}]$ case. Second, in the latter, non-self-consistent case, the gap edges recognized by G_{LDA} and $\text{Im}\Sigma_{GW}\,[G_{LDA}]$ are different. This mismatch in the gap edge, Δ, exaggerates the FWHM according to $\sim (\varepsilon + \Delta)^2 - \varepsilon^2 = 2\Delta \cdot \varepsilon + \Delta^2$, where ε is the QP energy measured from each gap edge, and $\Delta \sim$[$(\mu - \mu_0)$ plus the shift of the respective edge relative to LDA] [20]. (N.B.: the QP states obtained from $\Sigma_{GW}\,[G_{LDA}]$ have an unphysical finite lifetime, $\sim \Delta^2$, at the gap edges!)

From Table I it also follows that PS-based GW schemes carry a built-in error *which is comparable with the LDA gap* —cf. the representative PS-based result (1.29 eV [2]; 7^{th} row) with our corresponding all-electron result (0.85 eV; 4^{th} row). In addition to the non-local many-body effect of the core elucidated above, there is the impact on the matrix elements of Σ_{GW} of the removal of the oscillations of the valence wave functions at the atomic sites [20]. Interestingly, the projector-augmented-wave result of Arnaud and Alouani [10] (6^{th} row) provides an independent measure of the error due to the latter source. We stress that, at the $\Sigma_{GW}\,[G_{LDA}]$ level, the uncontrolled PS approximation [2, 3, 4, 5, 6, 7, 8] is masked by the neglect of self-consistency —which introduces an opposite-sign error, and results in apparent agreement with the experimental gap [27].

Ge is isoelectronic with Si, and has the same diamond crystal structure. It is then reassuring that an analysis along the above lines confirms that our conclusions concerning the impact of the deep core states and many-body self-consistency apply in the case of Ge as well. Note, in particular, the 1^{st} column of Table II: The additional opening of the *indirect* gap at L obtained from $\Sigma_{GW}\,[G_{GW}]$ (3^{rd} row) —relative to the value obtained

TABLE II: QP band gap and occupied bandwidth of Ge. In the 5^{th} row we exclude the contribution from the $3d$ states to the valence-state self-energy. For other conventions, see Table I; for the 6^{th} row, see text.

	absolute gap: ΓL	Direct gap at Γ	Indirect gap: ΓX	Occupied Bandwidth
Experiment [24]	0.74	0.89	. . .	. . .
LDA (FLAPW)	0.35	-0.20	0.66	12.82
$\Sigma_{GW}\,[G_{GW}]$	0.79	1.51	0.71	14.77
$\Sigma_{GW}\,[G_{LDA}]$	0.51	1.11	0.49	13.12
$\Sigma_{GW}\,[G_{LDA}]$, no $3d$'s	0.51	1.11	0.49	13.12
PS-based +CPP [6]	0.73	0.85	1.09	. . .

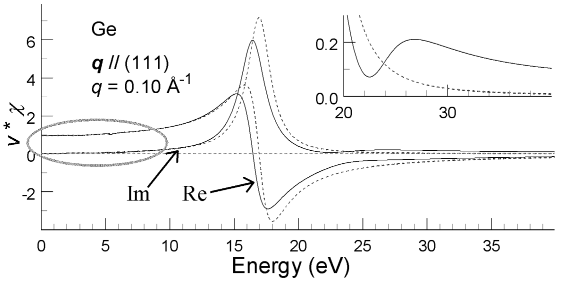


FIG. 2: Solid lines: Real and imaginary parts of the density-response function of Ge, obtained in an adiabatic, local, implementation of time-dependent density-functional theory [23]. The dashes correspond to the elimination of transitions from the $3d$ states. The $3d$-onset is highlighted in the inset.

from $\Sigma_{GW}\,[G_{LDA}]$; 4^{th} row— leads to excellent agreement with experiment [24].

A significant issue in Ge is the role of the $3d$ semicore states (which lie ∼25eV below the gap). In fact, in the standard PS-based GW approach, the correct topology of the QP band structure along ΓL *—which is automatically produced in our results—* is obtained only upon introducing the $3d$'s in the gap problem via a phenomenological core-polarization potential (CPP) model [6]. Our all-electron approach places us in a position to address this issue. Shown in Fig. 2 is the density-response function [23] of Ge for a small wave vector, evaluated with and without the contribution from the $3d$'s. Evidently, the screening is virtually identical in both calculations for energies up to ∼10 eV, by virtue of the weak transition strength of the $3d$'s. Thus, the impact of the CPP model [6] is not physically justified. Moreover, if the $3d$'s are excluded from *both* W and G, our all-electron QP band structure remains unchanged (5^{th} row). We conclude that *the 3d's play no role within the GWA.*

Now, although our calculated indirect gap at L agrees with experiment very well, other aspects of our GW results for Ge are less satisfactory. As illustrated in Table II (3^{rd} column), the empty states near X lie slightly below the lowest empty state at L —contrary to experiment [24]. This ordering, which is reversed in the absence of screening —i.e., if Σ is evaluated within HF [20]— provides a signature of the limitations of the GWA (note also our results for the Ge direct gap and the Si and Ge bandwidths). Thus, we expect that the localized Ge $3d$ states may induce short-range correlation effects in Σ; within the conserving method, such effects are to be included by adding an appropriate Φ-functional to Eq. (1) [29].

In summary, our results embody a new paradigm for *ab initio* QP theory, as we have demonstrated the non-trivial role played by the deep core states, and many-body self-consistency, in the QP gap problem; self-consistency —and thus, the fulfillment of the conservation laws— was also shown to impact the QP lifetimes. These effects, whose inclusion leads to excellent agreement with experiment for the Si absolute gap, and for the Ge indirect gap al L, are masked in the standard GW schemes. The Ge $3d$ semicore states were found to play no role in the GW gap; their impact is likely to come via effects beyond the GWA. Our results indicate that a fundamental description of the entire valence QP band structure of Si and Ge within 0.1eV requires the inclusion of mechanisms beyond the GWA.

This work was supported by NSF and NERSC.

[*] Present address: Department of Physics, University of California, Davis, California 95616.

[1] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[2] M. Hybertsen and S. G. Louie, **55**, 1418 (1985).
[3] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. Lett. **56**, 2415 (1986).
[4] L. Hedin, Phys. Rev. **139**, A796 (1965).
[5] For a recent review, see, e.g., W. G. Aulbur, L. Jönsson, and J. W. Wilkins, in *Solid State Physics,* edited by H. Ehrenreich (Academic, Orlando, 1999) Vol. 54, p.1.
[6] E. L. Shirley, X. Zhu, and S. G. Louie, Phys. Rev. Lett. **69**, 2955 (1992); Phys. Rev. B **56**, 6648 (1997).
[7] M. Rohlfing, P. Krüger, and J. Pollmann, Phys. Rev. B **57**, 6485 (1998).
[8] H. N. Rojas, R. W. Godby and R. J. Needs, Phys. Rev. Lett. **74**, 1827 (1995).
[9] N. Hamada, M. Hwang, and A. J. Freeman, Phys. Rev. B **41**, 3620 (1990).
[10] B. Arnaud and M. Alouani, Phys. Rev. B **62**, 4464 (2000).
[11] G. Baym and L. P. Kadanoff, Phys. Rev. **124**, 287 (1961); G. Baym, Phys. Rev. **127**, 1391 (1962).
[12] A. Schindlmayr, Phys. Rev. B. **56**, 3528 (1997).
[13] S. V. Faleev and M. I. Stockman, Phys. Rev. B 62, 16707 (2000).
[14] H. J. de Groot, P. A. Bobbert, and W. van Haeringen, Phys. Rev. B **52**, 11000 (1995).
[15] B. Holm and U. von Barth, Phys. Rev. B **57**, 2108 (1998).
[16] W.-D. Schöne and A. G. Eguiluz, Phys. Rev. Lett. **81**, 1662 (1998).
[17] B. Holm, Phys. Rev. Lett. **83**, 788 (1999).
[18] H. J. de Groot *et al.*, Phys. Rev. B **54**, 2374 (1996); P. García-González and R. W. Godby, *ibid.* **63**, 075112 (2000).
[19] J. M. Luttinger, and J. C. Ward, Phys. Rev. **118**, 1417 (1960).
[20] Wei Ku, Ph. D. Thesis, The University of Tennessee (2000); Wei Ku, and A. G. Eguiluz, to be published.
[21] P. Blaha, K. Schwarz, and J. Luitz, *WIEN97* (Techn. Universität Wien, Austria, 1999).
[22] Here the "j-diagonal" approximation is employed as usual.
[23] Wei Ku and A. G. Eguiluz, Phys. Rev. Lett. **82**, 2350 (1999).
[24] *Zahlenwerte und Funktionen aus Naturwissenschaften und Technik*, Landolt-Bornstein, New Series, Vol. III, Pt. 17a (Springer, New York, 1982).
[25] The corresponding contribution of the deep core states

via the higher-order terms in Eq. (1) is negligible.

[26] The relativistic 2P ($\kappa = -2, +1$) states contribute more than 90% to the opening of the Si gap, as they extend more in space, and have similar angular momentum character as the states that define the gap [20].

[27] The exaggerated opening of the gap obtained in a previous self-consistent GW calculation [16] is a consequence of the use of the PS approximation [20].

[28] P. M. Echenique *et al.*, Chem. Phys. **251**, 1 (2000).

[29] For a pedagogical discussion of the conservation laws, see A. G. Eguiluz and Wei Ku, in *Electron Correlations and Materials Properties*, edited by A. Gonis, N. Kioussis, and M. Ciftan (Kluwer-Plenum, N. Y., 1999), p. 329.